Bayesian Model for Sources Intensities

Lazhi Wang

Statistics Department, Harvard University

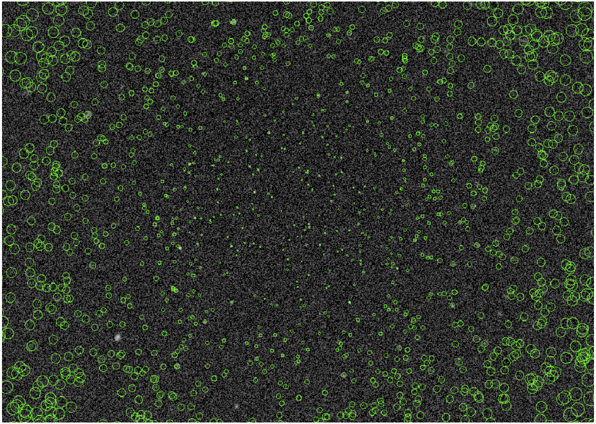
March 3, 2015

Outline

1. Background and goals of the project
2. Hierarchical Bayesian model
3. Frequency properties of the model via extensive simulation studies
4. Testing the existence of dark sources:
    - Calculation of test-statistic and posterior predictive p-value
    - Frequency properties of the *ppp* via simulation study
5. Real Data Application

Data

- Y_i, background contaminated photon count in a source region over a period of time $\mathcal{T}$.
- X, photon count in the exposure of pure background over $\mathcal{T}$.



Goals of the Project

1. To develop a fully Bayesian model to infer the distribution of the brightness (luminosity function) of all the sources in a population.

Goals of the Project

1. To develop a fully Bayesian model to infer the distribution of the brightness (luminosity function) of all the sources in a population.

2. To identify the existence of "X-ray" dark sources in the population.
    - "X-ray" dark sources: sources that do not generate X-rays.

Basic Hierarchical Bayesian Model

- Level I:

$$\begin{aligned} Y_i &= \mathcal{S}_i + \mathcal{B}_i \\ \mathcal{S}_i \big| \lambda_i &\sim \text{Poisson}(r_i e_i \mathcal{T} \lambda_i) \\ \mathcal{B}_i \big| \xi &\sim \text{Poisson}(a_i \mathcal{T} \xi) \\ X \big| \xi &\sim \text{Poisson}(A_b \mathcal{T} \xi) \end{aligned}$$

  - $\mathcal{S}_i$ (counts): number of photons from source i in the source region,
  - $\mathcal{B}_i$ (counts): number of photons from the background in the source region,
  - λ_i (counts/s/cm^2): the intensity of source i,
  - ξ (counts/s/pixels): the intensity of background,
  - t (seconds): exposure time,
  - e_i (cm^2): the telescope effective area,
  - r_i:proportion of photons from source i expected to fall in source region,
  - a_i (pixels): the size of source region i,
  - A_b (pixels): the size of background region.

$\mathcal{S}_i$, $\mathcal{B}_i$, λ_i, ξ are all unobserved/latent, t, e_i, r_i, a_i, A_b are all known constant. Y_i, X are observed data.

Basic Hierarchical Bayesian Model

- Level II:

$$\xi \sim \text{Gamma}[\mu_0, \theta_0]$$

$$\lambda_i | \mu, \theta, \pi_d \begin{cases} = 0 & \text{with probability } \pi_d, \\ \sim \text{Gamma}[\mu, \theta] & \text{with probability } 1 - \pi_d. \end{cases}$$

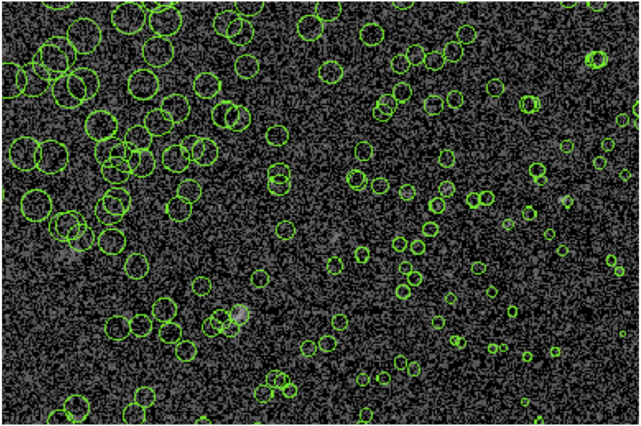
- Level III: Prior on the hyper-parameters π_d, μ, θ

$$\pi_d \sim Unif(0, 1)$$

$$P(\mu, \theta) \propto \frac{1}{c_1^2 + (\mu - c_2)^2} \frac{1}{c_3^2 + (\theta - c_4)^2} I_{\mu>0,\theta>0},$$

Model Extension I: Overlapping Sources

- Some source regions overlap.



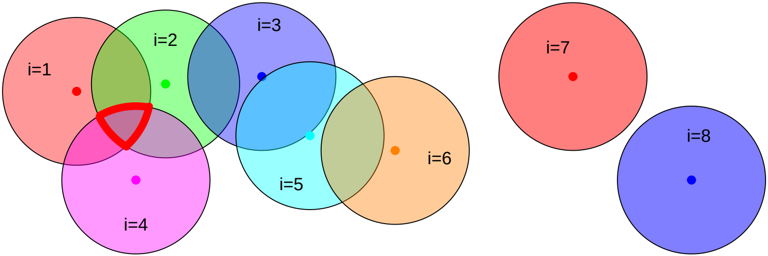
Model Extension I: Overlapping Sources

- Notation:
  - s is the set of indices of source regions that defines the segment. For example, the highlighted segment is $s = \{1, 2, 4\}$.
- Level I model:

$$Y_s = \mathcal{S}_s + \mathcal{B}_s = \sum_{i \in s} \mathcal{S}_{s,i} + \mathcal{B}_s,$$

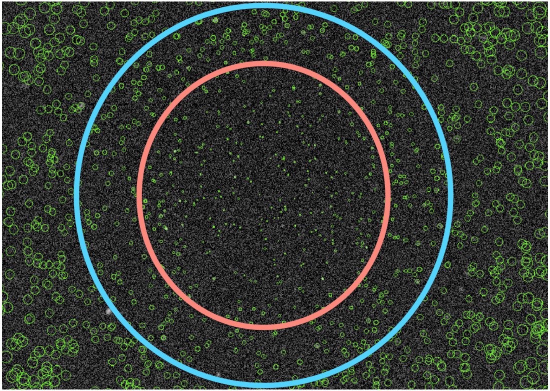
$$\mathcal{S}_{s,i} \big| \lambda_i \sim \text{Poisson}(r_{s,i} e_s \mathcal{T} \lambda_i)$$

$$\mathcal{B}_s \big| \xi \sim \text{Poisson}(a_s \mathcal{T} \xi)$$



- In our data, the background intensity has an increasing trend from the center to the edge of the telescope.

Projected Angle (arcmin)	0-6	6-8	8-16
Intensity (counts/pixels)	0.0010	0.0104	0.0108



Model Extension II: Different Background Intensities

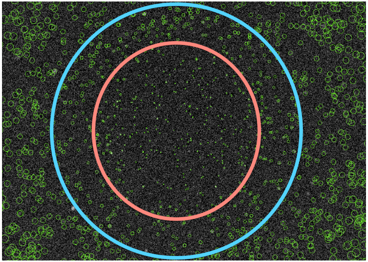
- Notation:
  - X_k (counts): number of photons collected in background region k over T seconds
  - ξ_k (counts/s/pixels): the background intensity in regions k
  - A_k (pixels): the size of background region k
  - $\mathcal{R}_k$: the collection of source segments in the background region k

- Model:
  - Counts in the pure background:

$$X_k \big| \xi_k \sim \text{Poisson}(A_k T \xi_k)$$

  - Counts in the source region $s \in \mathcal{R}_k$:

$$B_s \big| \xi_k \sim \text{Poisson}(a_s T \xi_k)$$



Simulation Setting

- Simulation Settings:

$$Y_i \sim \text{Poisson}(\lambda^* + \xi^*), \text{ for } i = 1, \cdots, 1000$$

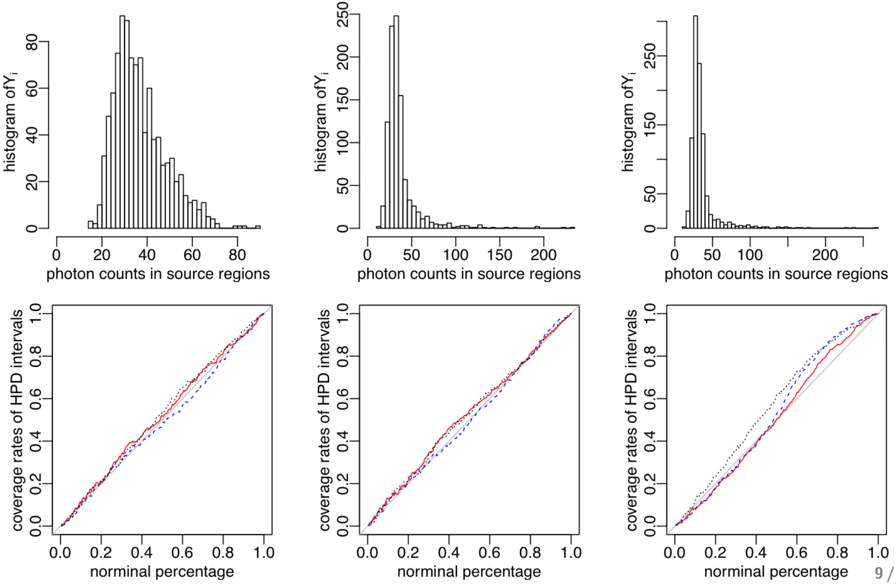
$$\lambda^* \begin{cases} = 0 & \text{with probability } \pi_d, \\ \sim \text{Gamma}[\mu^* = 15, \theta^*] & \text{with probability } 1 - \pi_d. \end{cases}$$

$$X \sim \text{Poisson}(2.5 \times 10^5),$$

- θ^*, π_d, ξ^* vary at different values:
  - ξ^*: 15, 30
  - θ^*: 50, 100, 300, 500, 1000
  - π_d : $0, 0.1, \cdots, 0.9$
- No overlapping sources
- Homogeneous background

Coverage Rates of 95% HPD Intervals

- $\pi_d = 0.5, \xi^* = 30, \mu^* = 15, \theta^* = 100, 500$ and 1000.



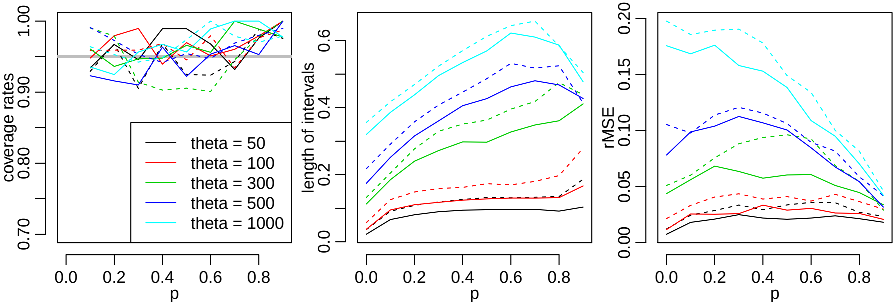
PME and HPD Intervals Estimates of π_d

- 100 replicate datasets for each simulation configuration.
- In each cell, the three summaries are (i) coverage rate of 95% HPD intervals, (ii) average length of intervals, (iii) root MSE

ξ^*	θ^*	π_d									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
15	50	–	93.4%	96.7%	94.6%	98.9%	98.9%	96.8%	93.2%	97.6%	100%
		0.02	0.07	0.08	0.09	0.09	0.1	0.1	0.1	0.09	0.1
		0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	100	–	94.8%	97.9%	99%	93.9%	97%	95.1%	96%	97.9%	100%
		0.04	0.09	0.11	0.12	0.12	0.13	0.13	0.13	0.13	0.17
		0.01	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02
	300	–	96.1%	93.6%	94.7%	94.8%	96.6%	95.6%	100%	98.9%	97.8%
		0.11	0.19	0.24	0.27	0.3	0.3	0.33	0.35	0.36	0.41
		0.04	0.06	0.07	0.06	0.06	0.06	0.06	0.05	0.04	0.03
	500	–	92.3%	91.6%	91%	96.3%	92.2%	95.3%	96.6%	95.3%	100%
		0.17	0.25	0.32	0.36	0.41	0.43	0.46	0.48	0.47	0.43
		0.08	0.1	0.1	0.11	0.11	0.1	0.08	0.07	0.05	0.03
	1000	–	93.5%	92.5%	95.7%	96.7%	95.7%	98.9%	100%	100%	97.8%
		0.32	0.39	0.44	0.5	0.53	0.57	0.62	0.61	0.59	0.48
		0.18	0.17	0.18	0.16	0.15	0.14	0.11	0.09	0.07	0.04

PME and HPD Intervals Estimates of π_d

- $\xi^* = 15$ (solid lines); $\xi^* = 30$ (dashed lines)



Hypothesis Testing for the Existence of Dark Sources

- Hypothesis Testing:

$$H_0 : \pi_d = 0, \quad H_a : \pi_d > 0.$$

- Reject H_0 if the p-value is low,

$$\text{p-value } = P(T(\mathbb{D}) > T^{obs} \big| H_0),$$

where $\mathbb{D} \sim H_0$ and $T(\mathbb{D})$ is a test statistic.

Hypothesis Testing for the Existence of Dark Sources

- Hypothesis Testing:

$$H_0 : \pi_d = 0, \quad H_a : \pi_d > 0.$$

- Reject H_0 if the p-value is low,

$$\text{p-value} = P(T(\mathbb{D}) > T^{obs} \big| H_0),$$

where $\mathbb{D} \sim H_0$ and $T(\mathbb{D})$ is a test statistic.

- However, $\mathbb{D} \big| H_0$ is unknown because μ and θ are unknown:

$$\lambda_i | \mu, \theta, H_0 \sim Gamma[\mu, \theta].$$

Hypothesis Testing for the Existence of Dark Sources

- Hypothesis Testing:

$$H_0 : \pi_d = 0, \quad H_a : \pi_d > 0.$$

- Reject H_0 if the p-value is low,

$$\text{p-value } = P(T(\mathbb{D}) > T^{obs} \big| H_0),$$

where $\mathbb{D} \sim H_0$ and $T(\mathbb{D})$ is a test statistic.

- However, $\mathbb{D} \big| H_0$ is unknown because μ and θ are unknown:

$$\lambda_i | \mu, \theta, H_0 \sim Gamma[\mu, \theta].$$

- Posterior predictive p-value (*ppp*):

$$ppp = P(T(\mathbb{D}) > T^{obs} \big| \mathcal{D}^{obs})$$

$$= \int P(T(\mathbb{D}) > T^{obs} \big| \mu, \theta, \pi_d = 0) P(\mu, \theta | \mathcal{D}^{obs}, \pi_d = 0) \mathrm{d}\mu \mathrm{d}\theta.$$

Hypothesis Testing for Existence of Dark Sources

- Estimation of *ppp*:

1. Draw $(\mu^{(t)}, \theta^{(t)})$ from $P(\mu, \theta \big| \mathcal{D}^{obs}, \pi_d = 0)$ for $t = 1, 2, \cdots, m$,
2. For each pair $(\mu^{(t)}, \theta^{(t)})$, simulate $\mathbb{D}^{(t)}$ from the null model and calculate $T^{(t)} = T(\mathbb{D}^{(t)})$,
3. Estimate *ppp* by

$$ppp \approx \frac{1}{m} \sum_{t=1}^{m} I\left(T^{(t)} > T^{obs}\right).$$

Hypothesis Testing for Existence of Dark Sources

- Estimation of *ppp*:

  1. Draw $(\mu^{(t)}, \theta^{(t)})$ from $P(\mu, \theta \big| \mathcal{D}^{obs}, \pi_d = 0)$ for $t = 1, 2, \cdots, m$,
  2. For each pair $(\mu^{(t)}, \theta^{(t)})$, simulate $\mathbb{D}^{(t)}$ from the null model and calculate $T^{(t)} = T(\mathbb{D}^{(t)})$,
  3. Estimate *ppp* by
  $$ppp \approx \frac{1}{m} \sum_{t=1}^{m} I\left(T^{(t)} > T^{obs}\right).$$

- Likelihood Ratio Test Statistics:

$$R(\mathbb{D}) = \frac{\sup_{\mu,\theta,\pi_d} L_a(\mu, \theta, \pi_d \big| \mathbb{D})}{\sup_{\mu,\theta} L_0(\mu, \theta \big| \mathbb{D})},$$

We use $T(\mathbb{D}) = log(R(\mathbb{D}))$ as the test statistic.

Two simplifications for the LRT:

- To obtain the likelihood $L_a(\mu, \theta, \pi_d \big| \mathbb{D})$ or $L_0(\mu, \theta \big| \mathbb{D})$, we need to integrate out all other parameters.

$$P_a(\mathbb{D} \big| \mu, \theta, \pi_d) = \int P(\mathbb{D} \big| \boldsymbol{\xi}, \boldsymbol{\lambda}) P(\boldsymbol{\xi}) P_a(\boldsymbol{\lambda} \big| \mu, \theta, \pi_d) d\boldsymbol{\lambda} d\boldsymbol{\xi}.$$

- No close form likelihoods if some source regions overlap and $\boldsymbol{\xi}$ is random.

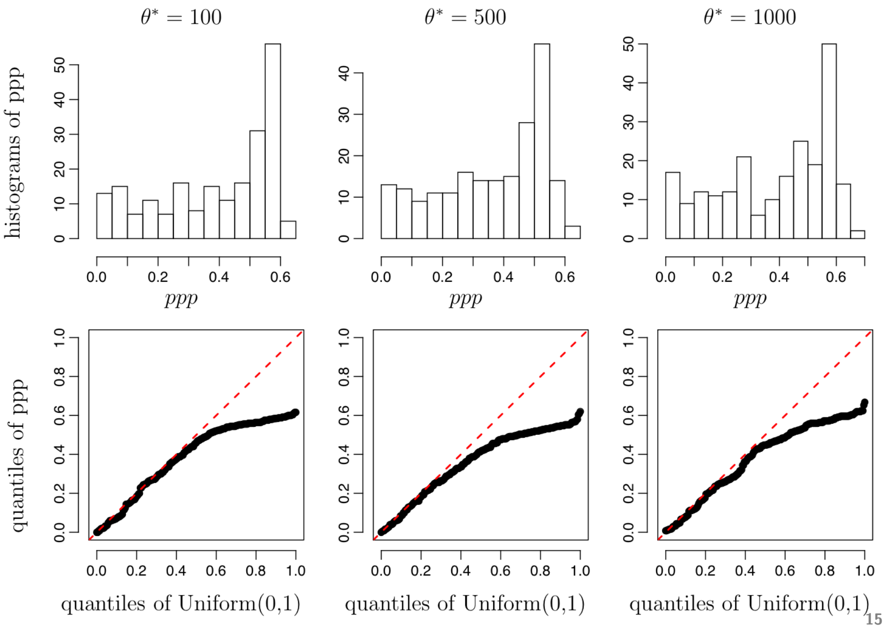
Two simplifications for the LRT:

- To obtain the likelihood $L_a(\mu, \theta, \pi_d \big| \mathbb{D})$ or $L_0(\mu, \theta \big| \mathbb{D})$, we need to integrate out all other parameters.

$$P_a(\mathbb{D} \big| \mu, \theta, \pi_d) = \int P(\mathbb{D} \big| \boldsymbol{\xi}, \boldsymbol{\lambda}) P(\boldsymbol{\xi}) P_a(\boldsymbol{\lambda} \big| \mu, \theta, \pi_d) d\boldsymbol{\lambda} d\boldsymbol{\xi}.$$

- No close form likelihoods if some source regions overlap and $\boldsymbol{\xi}$ is random.
- Two simplifications in the calculation of likelihoods:
  1. Simplification 1: Plug in $A_k \hat{\xi}_k t = X_k$.
     - Hardly changes the posterior distribution of hyper-parameters!
  2. Simplification 2: Likelihoods are calculated based on non-overlapping sources $\mathbb{D}^*$: $L_a(\mu, \theta, \pi_d \big| \mathbb{D}^*)$ and $L_0(\mu, \theta \big| \mathbb{D}^*)$
- $T(\mathbb{D}^*) = log(R(\mathbb{D}^*))$ is still a valid statistic.

Simulation Study: Distribution of *ppp* under H_0



Simulation Study: Power of the Test

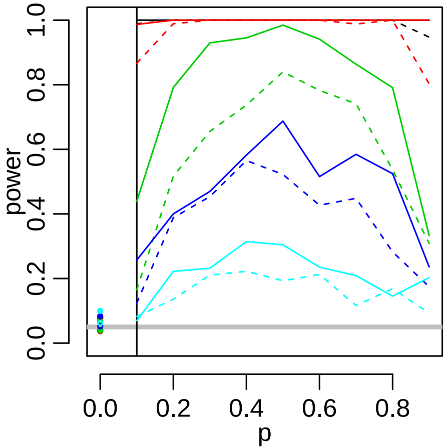
Table 3: The rejection rates of our hypothesis testing procedure.

ξ^*	θ^*	π_d									
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
15	50	6.8%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	100	3.7%	98.7%	100%	100%	100%	100%	100%	100%	100%	100%
	300	3.9%	43.9%	79.2%	93%	94.5%	98.5%	94.1%	86.4%	79.1%	33.3%
	500	6.3%	25.7%	40%	47%	58.2%	68.8%	51.6%	58.5%	52.5%	23.6%
	1000	6.1%	6.9%	22.2%	23.2%	31.4%	30.4%	23.5%	20.9%	14.5%	20.3%
30	50	4.1%	98.9%	100%	100%	100%	100%	100%	100%	100%	94.7%
	100	7.8%	86.7%	98.9%	100%	100%	100%	100%	98.8%	100%	80.3%
	300	6.9%	16.5%	51.7%	65.6%	73.7%	84%	78.3%	74.1%	53.7%	30.9%
	500	5.2%	12.4%	38.8%	45.4%	56.5%	52.2%	42.7%	44.8%	28.3%	17.4%
	1000	6.2%	8.3%	13.6%	21.1%	22.2%	19.3%	21.2%	11.6%	16.9%	9.3%

* Based on 100 replications.

Simulation Study: Power of the Test

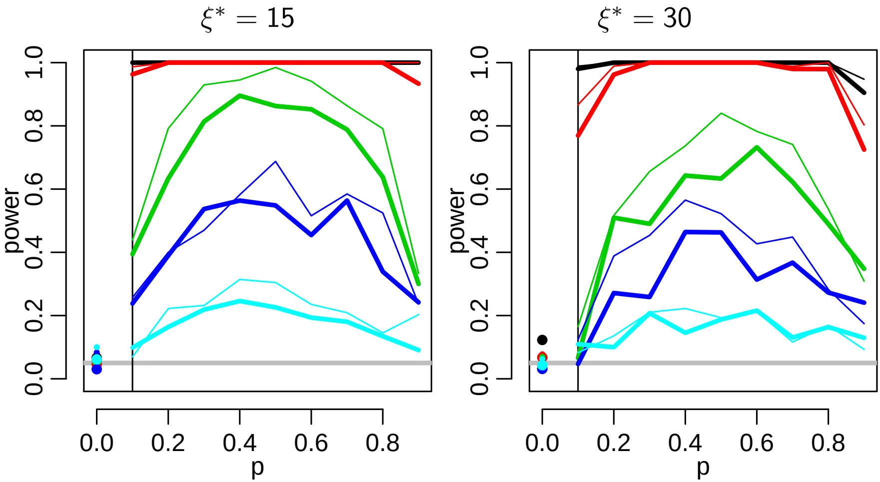
- $\xi^* = 15$ (solid lines); $\xi^* = 30$ (dashed lines)



* Based on 100 replications.

Simulation Study: Power of the Test

- thin lines: all the data are used to calculate the test statistic
- thick lines: 80% of the data are used to calculate T.



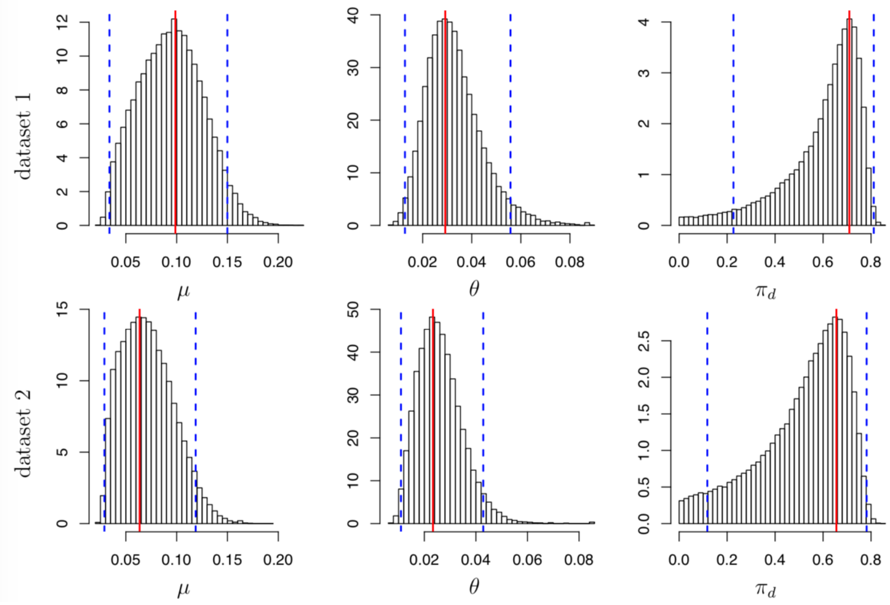
* Based on 100 replications.

Real Data: subsets of the Chandra/HRC-I observation of the stellar open cluster, NGC 2516.

- Dataset 1:
  - 649 sources within 6 arcmin from the center of the field
  - 525 non-overlapping sources
  - average source regions $\approx$ 1400 pixels
  - background is assumed to be spatially uniform
- Dataset 2:
  - 1169 sources within 8 arcmin from the center of the field
  - 747 non-overlapping sources
  - average source regions $\approx$ 3847 pixels
  - background is assumed to be piecewise uniform ($<$6 and 6-8 arcmin)

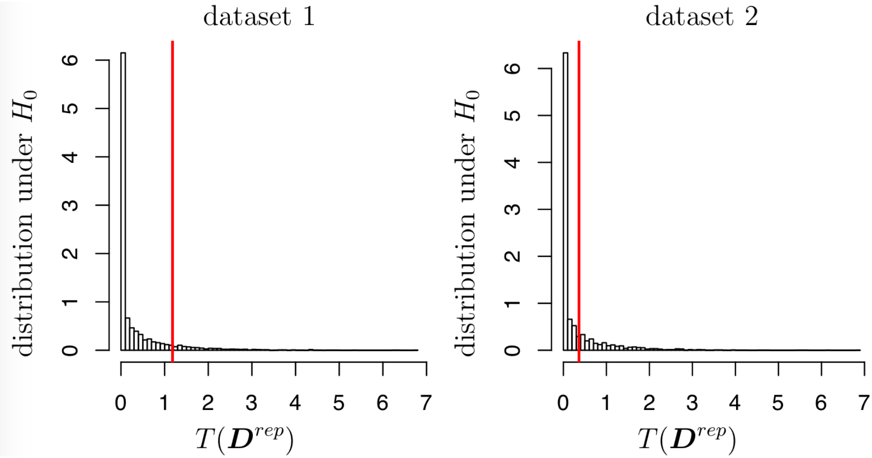
  - data between 6-8 arcmin from the center of the field:
    - 520 source
    - 227 non-overlapping sources
    - average source regions $\approx$ 6900 pixels

Real Data Analysis

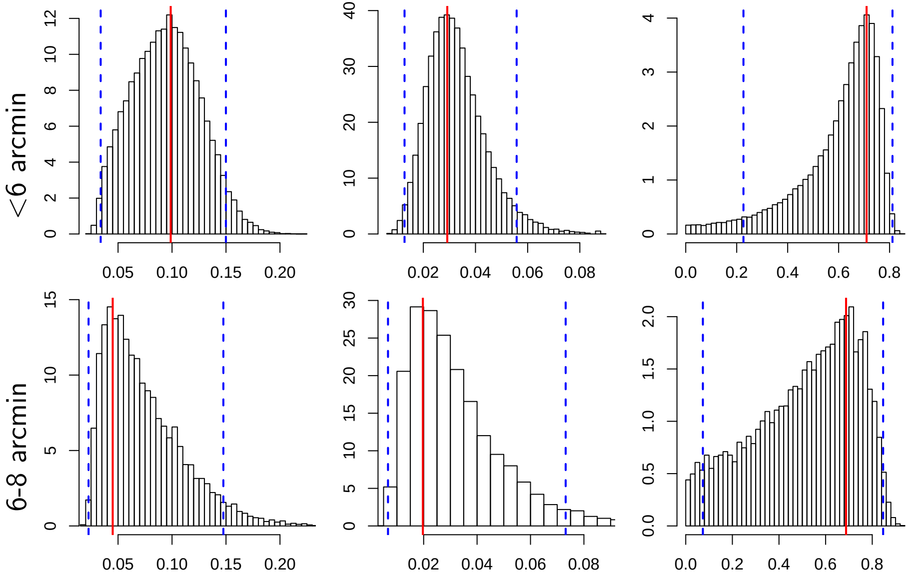


Real Data Analysis

- Dataset 1: $T(D^{obs}) = 1.181$ and $ppp \approx 8.9\%$.
- Dataset 2: $T(D^{obs}) = 0.363$ and $ppp \approx 23.2\%$.



Real Data Analysis



Real Data Analysis

- If we compute the likelihoods based on the 227 non-overlapping sources between 6-8 arcmin from the center of the field,

$$T^{obs} = 0.$$